Contract No.: SSA-0600-00-60021
MPR Reference No.: 8761-940

MATHEMATICA
Policy Research, Inc.

National Survey of SSI Children and Families User's Manual for Restricted and Public Use Files

Draft Report

April 2004

Jennifer Gillcrist
David Edson

Submitted to:

Social Security Administration
3532 Annex Building
6401 Security Boulevard
Baltimore, MD 21235
Telephone: (410) 965-5519

Project Officer:
Michele Adler

Submitted by:

Mathematica Policy Research, Inc.
600 Maryland Ave., SW, Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director:
Susan Mitchell

PREFACE

For more information on the National Survey of SSI Children and Families or to access files described herein, please contact Michele Adler at the Office of Disability and Income Security Programs, Social Security Administration, 3532 Annex Building, 6401 Security Boulevard, Baltimore, MD 21235 or at michele.c.adler@ssa.gov.

CONTENTS

Chapter **Page**

CONTENTS *(continued)*

Chapter **Page**

LIST OF TABLES

I. INTRODUCTION

The National Survey of SSI Children and Families (NSCF) collected data on children and young adults with special health care needs and their families who received or applied for Supplemental Security Income (SSI). The survey was sponsored by the Social Security Administration's (SSA) Office of Research, Evaluation, and Statistics until 2002 and thereafter the Office of Disability and Income Security Programs. The survey had two major objectives:

- To provide information on the characteristics, experiences, and needs of the current cross-section of SSI child recipients and their families
- To evaluate the effects of the Personal Responsibility and Work Opportunity Act of 1996 (P.L. 104-193, otherwise known as welfare reform) on SSI children and their families.

As the first national survey of SSI children since 1978, the NSCF provides information of substantial interest to SSA and policy analysts in other agencies and research institutions.

In 1999, SSA designed the NSCF with advice from an expert panel and technical assistance provided by MPR, the research firm that conducted the survey. The sample size was designed to provide reliable statistical estimates for a variety of analytic populations and to address policy questions of interest to SSA, such as assessing the effects of welfare reform on children in the SSI program, as well as addressing a variety of other issues of interest to SSA and others in the policy community. The questionnaire was designed to collect a rich array of data on children's health and socioeconomic status. By drawing on questions used in other national surveys on children's health and disability issues, the NSCF questionnaire yielded data for comparative analysis. Data collection began in July 2001 using computer-assisted telephone interviewing (CATI). Beginning in November 2001, in-person interviews were conducted with telephone nonrespondents using computer-assisted personal interviewing (CAPI). CATI and CAPI data collection was completed in August 2002. In all, respondents for 8,726 children and young adults

who have experience with the SSI program—either as current beneficiaries, former beneficiaries, or applicants who never received benefits—were interviewed. An additional 516 sample members were determined to be ineligible to participate in the survey. The survey was completed with a weighted response rate of 74.4 percent and an unweighted response rate of 77.2 percent.

The User's Manual provides information on the survey design, data collection and data preparation. It also describes the content and format of the restricted and public use data files and codebooks. Chapter II explains NSCF's two-stage probability sample design. Chapter III discusses the questionnaire's design, the incorporation of child and young adult versions of the questionnaire, question sources, and the 15 sections of the questionnaire (Sections A-O). Chapter IV describes the dual-mode (CATI and CAPI) data collection. Chapter V describes the creation of the restricted and public use files and associated data preparation. Chapter VI discusses the derivation of appropriate variance estimates. Finally, Chapter VII details the contents and layout of the restricted and public use files and codebooks. Chapter VII also discusses the proper use of the weighting variables.

II. SAMPLE DESIGN

Due to NSCF's multiple and competing objectives, MPR used a complex allocation algorithm to ensure adequate sample sizes for survey estimates for more than 100 analytic populations and subpopulations at a minimum expected cost. The analytic populations for current recipients are defined by age (under/over 17), gender, type of impairment (mental versus other), living situation, and duration of SSI receipt. For welfare reform analyses, the sample includes: (1) children subject to redetermination with SSI benefits continued, (2) children subject to redetermination with SSI benefits ceased, (3) children not subject to redetermination, and (4) children having previous contact with SSI, but not receiving SSI benefits at the time of welfare reform.

The sampling frame consists of children and young adults in the SSI applicant and beneficiary files at two time points: December 1996 and December 2000. MPR processed the 100 percent SSI extract files for these two time points and the "children's universe" file of children subject to redetermination as required by welfare reform. The December 1996 100 percent extract file contained 3,069,383 records and the December 2000 100 percent extract file contained 4,374,545 records. The children's universe file contained approximately 330,000 records.

The children eligible for the NSCF included all children that were recipients of SSI at the time of welfare reform or were recipients in December 2000. For this survey, children were classified as recipients if the current pay status information on the extract record was not a terminated status code. Children that were not recipients at either of these time points were also eligible for this study if the child either had been a recipient or applied for SSI, and the application date was after January 1, 1992.

Children that were recipients at the time of welfare reform were classified into sampling strata based on redetermination status (subject and not subject to redetermination) and the outcome of the redetermination process (continued on SSI or were denied SSI). For the analysis of welfare reform, particular interest is in the children that were subject to redetermination and separate sampling strata were formed for children subject to the redetermination process and continued on SSI or for children subject to the redetermination and were denied SSI. These two strata included all children meeting these criteria without regard to the child's age or current recipient status. Because of the issues related to the transitioning of children to the adult eligibility criteria, a separate stratum of children was formed that included SSI recipients that were either 17 or 18 years in December 1996 and were either on SSI at welfare reform and not subject to redetermination or not on SSI at welfare reform nor currently, but had previously received benefits or had applied after January 1, 1992.

For children that were current recipients (as of December 2000), three sampling strata were defined on the basis of whether the child was on SSI at welfare reform and not subject to redetermination or was not on SSI at welfare reform, and the age of the child. Children under 17 years were classified into two sampling strata. Once again, because of the issues related to the transitioning of children to the adult eligibility criteria, a separate stratum of young adults was formed that included current SSI recipients that were either 17 or 18 years and either were not on SSI at welfare reform or were SSI recipients at welfare reform but were not subject to redetermination (see Table II.1).

TABLE II.1
SAMPLING STRATA DEFINITIONS

Sampling Strata	Age	Sample
Total		**11,971**
1. Children and young adults who were SSI recipients at welfare reform and were subject to redetermination and were continued	All ages	2,377
2. Children and young adults who were SSI recipients at welfare reform and were subject to redetermination and were denied	All ages	2,438
3. Children and young adults who were SSI recipients at welfare reform and not subject to redetermination, but are not currently SSI recipients	Under 17 at welfare reform	1,059
4. Children and young adults who were not SSI recipients at welfare reform and are not currently SSI recipients	Under 17 at welfare reform	1,433
5. Young adults who are not currently SSI recipients and were either A. SSI recipients at welfare reform and not subject to redetermination B. Not SSI recipients at welfare reform	17 to 18 at welfare reform	935
6. Children who are currently SSI recipients and were SSI recipients at welfare reform, but not subject to redetermination	Under 17 At Survey	1,341
7. Children who are currently SSI recipients and were not SSI recipients at welfare reform	Under 17 at survey	1,381
8. Young adults who are currently SSI recipients and were either A. SSI recipients at welfare reform and not subject to redetermination B. Not SSI recipients at welfare reform	17 to 18 at survey	1,007

The NSCF used a two-stage probability sample design with the selection of primary sampling units (PSUs) that were formed using counts of children based on the SSI applicant and beneficiary files (described above) aggregated to single or multiple county-level units. PSUs, based on single or multiple adjacent counties, were constructed using SSI program files and

selected to form a nationally representative sample. The 74 PSUs selected contain more than 916,000 of the 3.5 million children in the survey population.

In the 74 sampled PSUs, the sample of children was allocated across eight sampling strata. These 592 allocations (74 PSUs x 8 sampling strata = 592 allocations) were then inflated to account for nonresponse and ineligible cases. The selection of the children was controlled by gender, age, presence of a mental disability, and geography. Initially, a larger sample of 27,465 children and young adults was selected and randomly partitioned into waves to control the sample release for reaching the target number of completed interviews. In total, a smaller sample of 11,971 cases was released for interviewing. For further information about the NSCF sample design, see Potter (2000).

III. QUESTIONNAIRE DESIGN

In designing the NSCF questionnaire, SSA was interested in obtaining answers to a number of questions regarding children and young adults with disabilities. The specific research questions were:

- What are the general characteristics of SSI children and their families (demographic, clinical, and family status)?
- What are the patterns of access to and utilization of health care among SSI children?
- What services do SSI children use?
- What are the costs associated with caring for a child with a disability?
- What is the impact on the family of having a child with a disability?
- What is the status of young adults with disabilities as they transition to adulthood?
- What is the impact of the 1996 welfare reform legislation on former child recipients in terms of their health, well-being, and transition to adult life?

A. QUESTIONNAIRE SECTIONS

The questionnaire incorporated two different versions: the child version and the young adult version. The versions were similar in content, but allowed for differences in living situations, SSI eligibility, and other age-specific issues between children and young adults.

The child version was designed for sample members who were under age 17 at the time of the survey. The young adult version was for sample members who were between 17 and 24 at the time of the survey. Both child and young adult questionnaire versions asked about the sample member's health status and functional limitations, health care utilization, health insurance coverage, receipt of services, and SSI experience. In addition, data were collected about the socioeconomic status of the sample members' households, including earned and unearned income, and housing characteristics. Both versions required about 70 minutes to administer. A

Spanish version of the questionnaire was also available in CATI and CAPI to ensure representation of Spanish-speaking families.

The questionnaire was divided into 15 sections, A-O, each of which addressed a particular topic, such as education and training. The nature of the questions including topics like household income and parental employment, as well as questions referring to the family's household circumstances at the time of welfare reform in 1996, made it necessary to identify a respondent to participate on behalf of a child sample member. Similarly, parent/guardian or proxy respondents were required for incarcerated sample members. Respondents for sample members under age 18 (non-incarcerated) were asked questions from all sections except Section N. Section N was a condensed version of the questionnaire that focused on questions that apply to imprisoned sample members. Sample members age 18 or older or their respondents answered questions from all sections except Sections J – Work and Childcare, and Section N – Imprisonment Module. Respondents for imprisoned sample members answered only Section A – Introduction and Screener, Section N – Imprisonment Module, and Section O – Closing and Interviewer Observations. The questionnaire sections are explained in more detail below:

Section A—Introduction and Screener—all sample members. This section included questions that identify and gain cooperation of the sample member and the respondent, and also included the household roster that collects information for all household members, including age, sex, and relationship to the sample member.

Section B—Disability Status and Functional Limitations—all sample members. This section screened sample members for the presence of a health condition, and then followed up with questions about the condition's nature, severity, and duration. The questions allow construction of disability indices by severity of reported limitations. Using several of the items together will allow classification of respondents into severity groups and facilitate comparisons

with other national data collections, such as the National Survey of Children with Special Health Care Needs (CSHCN), and the National Health Interview Survey (NHIS).

Section C—Health Care Utilization—all sample members. This section collected descriptive information on how frequently the sample members use doctors, hospitals, emergency room care, and prescription drugs. In addition, questions were asked about the family's out-of-pocket expenses for health care in the last twelve months, and the sample member's unmet health care needs.

Section D—Health Insurance—all sample members. Section D asked about the type of health insurance the sample member had (Medicaid, State Children's Health Insurance Program (SCHIP), employer or union, military, or directly from an insurance provider), who paid for the coverage, and about any episodes when the sample member was without health coverage.

Section E—Education and Training—all sample members. These questions collected data on the sample member's educational attainment, as well as receipt of special education, early intervention, and vocational education services. Because young adults may receive different kinds of training than children, questionnaire routing differed for child versus young adult respondents.

Section F—Programs and Services—all sample members. This section covered the programs and services used or needed by the families of SSI recipients, including therapy services and family-centered services such as respite care and family counseling. Section F also collected data on who paid for the services, unmet needs for services, and the out-of-pocket costs to the family.

Section G—Impact on Family and Self—all sample members. This section asked questions about quality of life issues such as food, housing, and monetary security. Items were included on

the child's behavior and social interactions, as well as how having a child with a disability impacted the family's interactions and living arrangements.

Section H—SSI Experience—all sample members. This section covers receipt of SSI benefits, and the family's experience with redetermination and the appeals process. Other items ask about how the family uses the SSI benefit. In addition, items asked about the family's familiarity with and use of a number of SSA-sponsored work incentive programs for SSI recipients, such as Plans for Achieving Self Support (PASS), Individual Development Accounts (IDA), and earned-income exclusions.

Section I—Employment—all sample members. Section I asked about the employment of sample members' parent(s)/guardian(s) and of young adult sample members themselves. In cases in which the sample member was married, information about the spouse's employment was also collected. Questions ask about the type of work performed, type of employer, hours worked, and wages earned. Questions also addressed how having a child with a disability affected parental labor force participation, and for young adults, their ability to work and their work experience.

Section J—Work and Child Care—sample members under age 18 only. Section J was asked only when the sample member was a child. It covered issues of the sample member's care while his or her parents are working or attending school. It also asked questions about who provided the childcare, the number of hours childcare is provided each week, the need for specialized childcare, satisfaction with childcare, and the cost of the care to the family. Parents or guardians of children who did not need childcare did not answer questions in this section.

Section K—Unearned Income and Assets—all sample members. This section included detailed questions on the family's receipt of unearned income including government benefits such as Food Stamps, Temporary Assistance for Needy Families (TANF), foster care payments, and unemployment compensation, and other unearned income, such as child support and pension

payments. Questions asked who in the household received the benefit or payment, and the amount received last month. Other questions asked about the value of the family's or young adult's assets at the end of the prior month and their overall debt burden.

Section L—Housing and Transportation—all sample members. Section L asked about the type of housing the sample member lives in, the cost of the housing, and the availability or need for modifications to accommodate persons with disabilities. Questions also asked about types of transportation used, and the sample member's need for special accommodations when using public transportation.

Section M—Background Information—all sample members. Section M collected demographic information about the sample member, the sample member's parents/guardians, and the sample member's spouse. The data collected include each individual's race, ethnic background, and education level as well as the language spoken in the household.

Section N—Imprisonment Module—imprisoned sample members only. Section N collected limited health and demographic information from a parent, guardian, or proxy of currently incarcerated sample members, whether young adult or child. For sample members who were incarcerated, only Section A, Section N, and the Section O (Closing Information) were asked. Incarcerated sample members were not interviewed. Parent/guardian or proxy respondents completed 191 interviews for imprisoned sample members.

Section O—Closing Information and Observations—all sample members. Section O was asked of all respondents and covered contact information for a possible future interview in two years. Section O also included interviewer observations.

B. QUESTIONNAIRE PATHING AND RESPONDENT TYPE

All sections of the NSCF questionnaire included distinct paths that depended on the sample member's age, living circumstances, and respondent type (RTYPE). At the beginning of the interview in Section A, the NSCF respondent was identified. For sample members under age 18, the respondent was always a parent or guardian (RTYPE=1). For sample members over age 18, the respondent could either be a parent/guardian, the sample member him or herself, or a proxy. A parent or guardian respondent was selected for sample members living at home if the sample member did not have spouse or child of their own (RTYPE=1). A parent or guardian respondent was selected for sample members over age 18 who were living at school (RTYPE=1). Sample members living independently (not in school) or having a family (spouse or child) of their own served as their own respondents (RTYPE=2). In cases in which the sample member could not complete the interview for him or herself due to a disability, a proxy was identified as the respondent (RTYPE=3).

Based on the information about the sample member's age and living situation, the sample member's respondent was designated to follow one of five major questionnaire paths: child path (CP), young adult parent path (YP), young adult path (YA), young adult proxy path (YX) or imprisonment path (JL). Table III.2 describes the characteristics that determined sample members' respondent type and questionnaire pathing.

TABLE III.1
QUESTIONNAIRE PATHING AND RESPONDENT TYPE

For Sample Member who is….	Respondent	Path
A child under 17	Parent/Guardian (RTYPE=1)	Child (CP)
A young adult 17 years of age	Parent/Guardian (RTYPE=1)	Young Adult Parent (YP)
A young adult (18^{+}) living at home (unmarried, no children)	Parent/Guardian (RTYPE=1)	Young Adult Parent (YP)
A young adult (18^{+}) living at school	Parent/Guardian (RTYPE=1)	Young Adult Parent (YP)
A young adult (18^{+}) living independently	Young Adult (RTYPE=2)	Young Adult (YA)
A young adult (18^{+}) living with parents plus his own spouse and/or child	Young Adult (RTYPE=2)	Young Adult (YA)
A young adult (18^{+}) living away from home and unable to respond	Proxy (a parent/guardian serving as proxy is considered a proxy) RTYPE=3	Young Adult Proxy (YX)
A child/young adult under 18 incarcerated in jail, prison or juvenile facility	Parent/Guardian (RTYPE=1)	Imprisonment Path (JL)
A young adult (18^{+}) incarcerated in jail or prison	Parent/Guardian (RTYPE=1) or Proxy (RTYPE=3)	Imprisonment Path (JL)

The respondent's path through the questionnaire also depended on the respondent's answers to individual questions. Not every respondent is asked every question in each section. Questions not asked of a particular respondent are designated as "legitimate missing" responses in the data files.

C. COMPARISON WITH OTHER QUESTIONNAIRES AND SURVEYS

When possible, questions were taken from past studies to allow for comparison with other datasets. Such comparisons were the primary focus of "Characteristics of the SSI Child Population: A Comparison Between the NCSF and Three National Surveys" (Ireys et al, 2004). Questions were created when they were unavailable in previous studies or when they were not

appropriate for the NSCF. Each question's source is included in the questionnaire and codebook documentation. The table below describes the sources from which NSCF questions were taken.

TABLE III.2
NSCF QUESTION SOURCES

Study	Study (Year)	Sponsoring Organization
Created	Created by MPR for NSCF	MPR
CSHCN	National Survey of Children with Special Health Care Needs (2000)	Maternal and Child Health Bureau (MCHB) and the National Center for Health Statistics (NCHS)
ICHP	Primary Care Assessment -- Children with Special Health Care Needs (1995)	Institute for Child Health Policy The University of Florida
FACCT	Screener to identify children with special health care needs (2000)	Foundation for Accountability
Mary Wagner	Contributed	SRI, International
MEPS	Medical Expenditure Panel Survey (2000)	Agency for Healthcare Research and Quality (AHRQ)
NEILS	National Early Intervention Longitudinal Study (1998)	U.S. Department of Education, Office of Special Education Programs
NHIS	National Health Interview Study (1999)	National Center for Health Statistics (NCHS)
NHIS-D	National Health Interview Study—Disability Supplement (1994)	National Center for Health Statistics (NCHS)
MPR	Contributed	MPR

NSAF	National Survey of America's Families (1999)	Numerous Foundations - (see http://www.urban.org/Content/Research/NewFederalism/NSAF/Overview/NSAFOverview.htm)
SIPP	Survey of Income and Program Participation (1996)	U.S. Census Bureau

IV. DATA COLLECTION

The NSCF was executed as a dual-mode survey—initial interview attempts were made using computer-assisted telephone interviewing (CATI) followed by computer-assisted personal interviewing (CAPI) of nonrespondents. The CAPI interviews were attempted with respondents who did not have telephones, were unlocatable via telephone attempts and electronic searches, or requested a telephone interview.

A. DATA COLLECTION PROCEDURES

The success of the NSCF data collection was determined by MPR's efforts to identify, locate and gain the cooperation of NSCF respondents. To gain cooperation and increase participation, MPR sent an advance letter to the parents (or representative payees) [1] of all sample members prior to the interview. The advance letter explained the purpose of the survey, offered assurances of confidentiality, and included a toll-free number for respondents to call with questions or to complete the interview at their convenience.

For cases with addresses that were no longer valid, about 70 percent of the sample, MPR used a variety of techniques for locating current addresses and telephone numbers, including searching commercially available databases, calling relatives and neighbors, and making in-person visits to the person's former neighborhood. Due to these extensive efforts, about 77 percent of cases with invalid addresses were located. In total, approximately 84 percent of the sample was located for interviewing.

[1] A representative payee is a person, agency, organization or institution selected to manage the SSI recipient's benefits when the recipient is under age 18 or is physically or mentally unable to do so himself.

An incentive payment experiment using checks, debit cards, and phone cards with a value of $10 was also incorporated into the data collection to encourage participation and show appreciation for responses. For more information on the implementation and results of the incentive payment experiment see Mitchell, et al. (2003).

B. THE PRETEST

A pretest conducted in July 2001 preceded the official commencement of CATI interviewing. The pretest consisted of 41 CATI interviews (23 child and 18 young adult) that were included in the final data file as part of the completed sample. The pretest identified minor changes to the CATI instrument, including the addition of some questionnaire probes. In addition, some minor programming problems were corrected for the full-scale CATI interviewing that began in August 2001

C. INTERVIEWING

The full-scale data collection effort began with the launch of CATI interviewing in August 2001. CAPI interviewing of telephone nonrespondents began in November 2001 and continued, concurrent with CATI interviewing, through July 2002. In total, 8,726 cases were completed—7,285 were completed via CATI and CAPI interviewers completed 1,441 cases in the field and supported the CATI effort by locating difficult to find respondents who subsequently completed their interviews via the telephone and were counted as CATI completes.

An additional 516 cases were determined to be ineligible based on survey criteria, which excluded deceased sample members, sample members no longer living in the continental United States or living in Medicaid facilities, and sample members identified as wards of the state. Final complete (n= 8,726) and final ineligible (n=516) cases are included on the restricted and public

use data files discussed in Chapters V and VII. Table IV.1 reports the final case disposition for all released cases in the sample.

TABLE IV.1
NSCF FINAL CASE DISPOSITION

Classification	Case Disposition	Cases	Percentage	Weighted Count	Weighted Percent
Total	**All attempted**	**11,971**	**100**	**3,502,650**	**100**
Respondents	**Total Respondents**	**9,242**	**77.21**	**2,604,344**	**74.36**
Eligible Complete	**Total Eligible Complete**	**8,726**	**72.90**	**2,469,553**	**70.51**
	Complete-CATI (FNL=10)[2]	6,350	53.05	1,776,114	50.70
	Complete-CATI phone in (FNL=14)	785	6.56	230,676	6.59
	Complete-CATI-SM incarcerated[3] (FNL=15)	150	1.26	30,428	0.87
	Complete field-CAPI (FNL=20)	1,400	11.69	423,396	12.09
	Complete field-CAPI-SM incarcerated (FNL=25)	41	0.34	8,939	0.26
Ineligible	**Total Ineligible**	**516**	**4.31**	**134,791**	**3.85**
	SM deceased (FNL= 440)	63	0.53	22,319	0.64
	SM in Medicaid institution (FNL= 450)	76	0.63	14,689	0.42
	SM moved outside continental U.S. (FNL= 460)	43	0.36	12,993	0.37
	Other ineligible (FNL = 490)	334	2.79	84,791	2.42
Located Nonrespondent	**Total Refusals**	**783**	**6.53**	**262,827**	**7.50**
	Refusal/refuse to communicate	102	0.85	38,457	1.10
	Refusal-unknown person	49	0.41	15,423	0.44
	Refusal-known respondent	311	2.60	105,896	3.01
	Refusal-no such person after breakoff	4	0.03	1,399	0.04
	Language barrier (non-Spanish)	17	0.14	6,498	0.19
	Language Spanish	6	0.05	2,845	0.08
	Physical/cognitive barrier	6	0.05	3,028	0.09
	No respondent available in field period	10	0.08	3,351	0.10
	Other eligible (SM known to be present)	276	2.30	85,292	2.43
	Effort ended/case retired	2	0.02	639	0.02
Unlocatable	**Total Non-Located**	**1,946**	**16.26**	**635,479**	**18.14**
	Unlocated by office	285	2.38	87,275	2.49
	Unlocated by field	1,657	13.85	547,477	15.63
	Max calls-no contact	4	0.03	727	0.02

[2] FNL is the final case status variable on the data files.

[3] SM denotes sample member

V. VARIABLE CONSTRUCTION AND EDITING

Cleaned, edited and coded survey data were used to create the NSCF survey databases. This chapter provides an overview of the variable construction, naming, and coding conventions that are used on the data files and accompanying codebooks.

A. PUBLIC USE VARIABLES

As noted earlier, the confidential nature of much of the data collected by the NSCF required that two versions of the database be developed. The NSCF Restricted Use File is the most comprehensive version of the NSCF survey database and is intended for internal SSA use. This version includes some confidential respondent information such as geographic data, SSA administrative data, and data about the specific imputation methods used.

The NSCF Public Use File was created for general use by researchers outside of SSA and was subjected to substantially more confidentiality masking procedures than the restricted version. The Public Use File was created by removing identifying variables, such as those pertaining to geography and SSA administrative variables. In addition to the removal of these variables, masking techniques, such as collapsing response categories into broader categories, were applied to other variables in order to protect respondent confidentiality. The techniques applied to create the public use variables are detailed in the public use codebook. A listing of the variables available on the NSCF Restricted Use (RUF) and Public Use (PUF) Data Files is included as Appendix A.

B. VARIABLE NAMING

The NSCF datasets contain administrative, questionnaire, and constructed variables. To aid in distinguishing between different types of data, the following variable-naming scheme was adopted:

- *Questionnaire variables* were collected directly from the respondent. These are indicated by section letter and question number, for example, C9. (For a copy of the questionnaire, contact Ms. Michele Adler at the Social Security Administration at michele.c.adler@ssa.gov.)
- *Administrative variables* were either drawn from the SSA administrative data or created by MPR. Those drawn from SSA administrative data retain their original names from the SSA administrative data file such as DIGDIB, the primary disability diagnosis code. Others were created by MPR for survey administration or analysis purposes. These variables include the weighting variables and other variables such as the unique identifier (MPRID) or the final case status variable (FNL).
- *Constructed variables* were created after data collection was completed using responses to questionnaire variables. Constructed variables are preceded with a "c_"—for instance, c_C9 or c_living_arrangements. These variables were developed to facilitate analytic use of the data.
- *Imputation flag variables* accompany questionnaire and constructed variables that have been imputed. The imputation flag variables indicate which responses were imputed and the method of imputation utilized for each imputed value. Imputation flag variables are preceded by an "i_"—for example, i_C9.
- *Public use variables* indicate that the variable has been masked for public use and replaced the original questionnaire variable on the public use file. Public use variables are preceded by a "p_"—for example, p_C9.
- *Public use imputation flag variables* are preceded by a "pi_"—for example, pi_B3. Unlike the imputation flag variables found on the restricted file, the public use imputation flag variables only indicate that the source variable was imputed, but not the method of imputation. The specific method of imputation (deductive, unweighted hot-deck, weighted hot-deck, or regression-based imputation) is suppressed to protect respondent confidentiality.

Table V.1 details the variable naming conventions, provides examples of each type of variable, and indicates whether a particular type of variable is included on the restricted use file, the public use file, or both.

TABLE V.1
VARIABLE NAMING CONVENTIONS

Naming Scheme	Type of Variable	Availability	Example
VARIABLE NAME	Questionnaire or administrative variable	Restricted and public use file	C9 or HUN
c_VARIABLE NAME	Constructed variable	Restricted and public use files	c_C9
i_VARIABLE NAME	imputation flag variable	Restricted use file only	i_C9
p_VARIABLE NAME	public use variable	Public use file only	p_C9
Pi_VARIABLE NAME	public use imputed flag variable	Public use file only	pi_C9

C. VALUE CODING CONVENTIONS

The following coding conventions are used on the files:

TABLE V.2
VALUE LABEL CONVENTIONS

Response Category		Type of Response
Numerical data	Character data	
.L	L	Legitimate missing—Due to questionnaire design, the respondent was not asked this question.
.D	D	Don't Know—Respondent answered "don't know"
.R	R	Refused—Respondent refused to answer question
.M	M	Missing data—Data are missing due to interviewer or programming error
.N	N	Not applicable
.I	I	Ineligible—Respondent was ineligible to complete the questionnaire, but was retained on the dataset for comparison purposes (n=516).

D. SELECTING POPULATIONS OF INTEREST

The NSCF questionnaire was designed to provide data on several distinct populations, including child recipients, young adult recipients, and those affected by welfare reform.[4] Researchers interested in specific populations will need to subset the data. Several subpopulations of interest are described in Table V.3 below:

TABLE V.3
SELECTING SUBPOPULATIONS OF INTEREST

Subpopulation of Interest	Select Final Status Codes – SAS Variable FNL	Select Other Variables	Subpopulation Counts
All completed cases	10, 14, 15, 20, 25	N/A	8,726
All young adult self-respondents	10, 14, 20	RTYPE=2	876
All proxy respondents	10, 14, 15, 20, 25	RTYPE=3	109
All non-incarcerated completed cases	10, 14, 20	N/A	8,535
All incarcerated completed cases	15, 25	N/A	191
All ineligible cases	440, 450, 460, 490	N/A	516
Current SSI recipients	10, 14, 15, 20, 25	c_ssi_last_month=1	4,935
Particular sampling strata	10, 14, 15, 20, 25	1-8, see strata definitions in Chapter II	See Table II.1

E. CODING OF OPEN-ENDED AND VERBATIM RESPONSES

The NSCF questionnaire included a number of questions that elicited open-ended responses that required coding. In order to facilitate analytic use of the data, these responses were grouped, or "coded". The methodology used to code each variable depended upon the variable's content.

Health Condition Coding

[4] A short list of useful variables and their values is included as Appendix B.

Information on the sample member's health conditions, either current or at the time of the sample member's SSI application, was recorded in Section B at questions B25, B37, B40, and in Section N for the incarcerated population at N18, N23, and N26. The respondent's verbatim responses were coded using the World Health Organization's International Classification of Disease–9th Edition (ICD-9) five-digit codes. Cases in which the respondent's answer did not provide sufficient specificity for coding to five digits were coded to three or four digits. Cases that lacked the specificity for even three- or four-digit codes were coded either to broader categories representing disease groups or coded as either physical, mental, or behavioral/emotional problems. Health conditions were coded to whatever level of specificity was provided for by the respondent's answers. In cases in which multiple, distinct conditions were recorded, the first three distinct conditions (or two conditions at questions B40 and N26) were recorded (for instance, three distinct conditions would be recorded at B40_1, B40_2, and B40_3). A quality assurance review of 10 percent of the health condition responses revealed a coding error rate of about 3.5 percent. This rate was higher than anticipated; therefore, a 100 percent review was initiated.

Following ICD-9 coding, the health condition variables were processed into sets of constructed variables that group health conditions into disease groups and converted the ICD-9 codes to four-digit SSA impairment codes. For respondent confidentiality purposes, the public use variable for the first health condition is masked to report whether the health condition was physical or mental. Variables for the second and third reported conditions were suppressed on the public use file.

Industry Coding

The industry that employed the sample member's parent(s)/guardian(s), the sample member, and the sample member spouse (as appropriate) was recorded in Section I. CATI and CAPI interviewers recorded the respondent's verbatim responses, which were then coded using the Census Bureau's North American Industry Classification System (NAICS) standard coding scheme.[5] A thorough quality assurance review of all industry coding was conducted.

Occupation Coding

The occupations of the sample member's parent(s)/guardian(s), the sample member, and the sample member spouse (as appropriate) were recorded in Section I. CATI and CAPI interviewers recorded the respondent's verbatim responses. These responses were coded using the Bureau of Labor Statistic's Standard Occupational Classification (SOC) scheme.[6] A thorough quality assurance review of all occupation coding was conducted.

Open-Ended Coding

Several questions on the NSCF questionnaire did not include any designated response categories, but were recorded strictly as verbatim responses. These "open-ended" responses were reviewed and response categories were created. Respondents' verbatim responses were then coded to the developed response categories.

Other Specify

A number of NSCF questions allowed for multiple responses as well as verbatim responses, which were recorded as "other specify." The "other specify" responses were reviewed after the survey administration and additional response categories were created as necessary. These post-

---

[5] The 2002 North American Industry Classification System edition codes were used. More information can be found at http://www.census.gov/epcd/www/naics.html.

[6] More information about the SOC codes can be found at http://stats.bls.gov/soc/socguide.htm#LINK2.

interview response categories do not appear in the NSCF questionnaire, but are included as response categories in the NSCF Restricted Use File Questionnaire Codebook (Appendix C).

The post-survey review also uncovered circumstances in which the interviewer did not properly code the response to one of the pre-designated survey response categories, but rather recorded the response in the "other specify" field. In these cases, edits were applied to correct the error. In some cases, this resulted in missing data because the interviewer error led to pathing that skipped an appropriate question and resulted in missing responses that were coded ".M."

F. ADDITIONAL CLEANING AND EDITING

NSCF data was thoroughly reviewed for discrepancies resulting from programming and interviewer error. In some circumstances, such as the "other specify" situation described above, post-survey edits were made to correct errors. For more information on data problems and the completeness of the survey database, see the Report on Data Quality in the National Survey of SSI Children and Families Database (Gillcrist et al. 2004).

G. CONSTRUCTED VARIABLES

The NSCF data file preparation included creating more than 500 constructed variables in order to simplify the NSCF data file and assist the user. The algorithms used to create the constructed variables are included in the datafile codebooks as SAS programming code. In many cases, the constructed variables replaced the original survey variables on the final data files. The majority of the constructed variables fall into one of the following categories:

Family and Living Situation Constructed Variables

The Family and Living Situation constructed variables were created from data collected in Section A of the questionnaire. In order to identify the correct NSCF respondent, a household roster (A41 and A92 question series) collected each household member's age, sex, and

relationship to the sample member. Each household member's information was recorded in a different position on the household roster (positions 2-14). The sample member's information was always recorded in position 1. In cases where the sample member was not the respondent, the respondent was asked to report information before that of anyone else in the household. Thus, the respondent's information should fill the second position. To aid the user, constructed variables were created that identified which position in the household roster, if any, the sample member's mother, father, spouse, and designated parent/guardian (1 and 2) occupy. Parent/guardian 1 and 2 were only designated for Child Path (CP)/Young Adult Parent Path (YP) cases, in which a parent or guardian was the respondent. A parent/guardian 2 was designated when the parent/guardian 1 (the respondent) reported having a spouse or partner. Other constructed variables identify the sample member's mother and/or father's age and type (biological/adoptive, foster, step or unmarried partner of parent).

Additionally, the constructed variables were created to describe the number of family members that reside in the household according to different definitions of family. These variables, c_family1, c_family2, and c_family3, provide household counts based on the number of individuals related by blood or marriage, foster relationships, and/or unmarried partners of parents.[7] Other constructed variable count the number of sample member's grandparents, children, and sample member children that reside in the household. Finally, a variable was constructed to describe the sample member's household composition (c_living_arrangements). This variable indicates if the sample member lives in a single parent (mother or father only) household, in a two-parent household, with other relatives, with a spouse, is incarcerated, etc.

[7] Detailed information about the construction of these variables can be found in the NSCF Restricted Use File Questionnaire codebook (Appendix C).

Health Condition Constructed Variables

The Health Condition constructed variables included developing variables that report the conditions in several different formats including: five-digit ICD-9 codes, four-digit SSA impairment codes, ICD-9 diagnosis groups, and SSA diagnosis groups.[8]

Logical Zero Constructed Variables

SSA requested the creation of "logical zero" constructed variables to assist the user with some statistical analysis packages by reducing the number of legitimate missing responses that originated from survey skip patterns. For example, if the respondent reported the sample member did not receive SSI last month (question H6), then the respondent would skip the follow-up question about how much SSI the sample member received last month (question H7). In general, when a respondent skips a question due to questionnaire logic, the recorded response in this circumstance was "legitimate missing" or ".L." However, the logical zero constructed variables were designed to "carry through" no or zero values to subsequent questions as appropriate. Thus, if the sample member reported not receiving SSI the previous month, then "logical zero" type constructed variable c_SSI_last_month_amt recorded the amount as "0." The underlying rationale was that if the sample member did not receive SSI last month, then the sample member received $0 in SSI benefits. Logical zero constructed variables and the "stem" question(s) that indicated the "no" or zero response carried through to the logical zero constructed variable are identified in the codebook user notes.

[8] Frequencies of the ICD-9 and SSA impairment codes as well as details for the diagnosis group constructions are included in the NSCF Restricted Use and Public Use codebooks.

Period/Amount Standardization Constructed Variables

Throughout the NSCF questionnaire, respondents had the option of reporting condition durations, income and expenditures for a variety of time frames, for instance, daily, weekly, monthly, etc. The NSCF questionnaire was designed with this flexibility with the expectation that allowing respondents to select the time frame (ideally, the time frame with which they were most comfortable) would improve data quality. In these situations, the amount and the period reported by the respondent existed as two distinct variables in the survey data. For example, at question F7 respondents could report having $1,200 in out of pocket expenses for physical, occupational, or speech therapy in the last twelve months or they could report $20 in out of pocket expenses last week. To aid the user, constructed variables were created to standardize the time frame associated with the variable, resulting in a single variable (i.e., c_F7) with one time frame in place of a pair of variables for the period and the amount (i.e., F7_AMT and F7_AN). This type of constructed variable was predominantly created for Section F questions regarding out of pocket expenses.

Pathing Combinations

The NSCF questionnaire design combined child and young adult versions, which resulted in identical questions being asked on multiple paths. When appropriate, a constructed variable was created that combined survey responses for all paths into one variable. For example, responses to question H6 (CP path) and H22 (YP, YA, YX paths) about the sample member's SSI status last month were combined to create constructed variable c_SSI_last_month. The constructed variable code included in the codebooks details the original questionnaire variables used to create the constructed variable.

Section K—Unearned Income Constructed Variables

The most extensive variable construction effort focused on Section K, which collected data on unearned income and assets for each member of the household. While many of the Section K constructed variables included the pathing combinations and logical zero changes previously described, Section K is unique in that the constructed variables often included more than one of those types of changes. In Section K, unearned income data was collected for each member of the household on two questionnaire pathings. Thus, many unearned income constructed variables involved both pathing combinations and logical zero changes. Indicator and aggregate variables were also constructed for unearned income data.

Section K was developed to include two separate paths for CP/YP respondents and for YA/YX respondents (see Chapter III, Section C for discussion of questionnaire paths), and the constructed variables combined the variables from the two paths when appropriate. Logical zero edits (described above) were also applied to many variables in this section, such as when either the entire household or a particular member of the household did not receive a benefit. In that case, a zero value appears for any household member not receiving the benefit. When there is no household member in a particular position (2-14), the value was set to legitimate missing (.L). For example, if the respondent reported that no one in the household received welfare benefits last month (K2 in the questionnaire), then the respondent skipped questions that asked who in the household received benefits (K3) and how much the benefits were (K5). Rather than reporting that the data was legitimately missing for each household member, the constructed variable (including the pathing combination) reports that the household member did not receive benefits (c_welfare_rcpt_1= NO (0)) and that the value of the benefits was zero (c_welfare_amt_1=0).[9]

[9] Receipt and amount variables were created for each position in the household roster (1-14). A 15th position is included for other members of the household whom the respondent did not initially report in the household grid, but

Indicator variables (such as c_welfare) were created for each of the unearned income categories to indicate whether anyone in the household received that particular benefit in the last month. Other constructed variables in Section K include variables that represent the each household member's total unearned income, for example c_unearned_income_1 through c_unearned_income_15.

H. IMPUTATIONS

In the NSCF, the data collection instruments were administered using computer-assisted interviewing technology, an approach that substantially reduces the extent of item nonresponse. However, some item nonresponse still existed. Item nonresponse included cases in which the question was not answered in error and cases where "don't know" or "refused" were recorded as responses. For the NSCF, several methods of imputation were used; the methods were selected based on the level of sophistication needed for the imputation and on the availability of data for the imputations. For some variables, two or more of these methods were used in combination to improve the imputations. After each imputation procedure, the imputed values were evaluated. If the initial imputed value was out of the acceptable range or inconsistent with other data for that case, the imputation was repeated until the imputed value was acceptable.

The four methods were:

1. Deductive (or logical) imputation

1. Unweighted hot-deck imputation

2. Weighted hot-deck imputation

3. Regression-based imputation

---

for whom receipt of benefits was reported later in the questionnaire. Thus, c_welfare_rcpt _1 and c_welfare_amt_1 refer to the person in the first position in the household roster, which is always the sample member.

As noted earlier, an imputation flag variable is associated with each variable in which an imputation has been made. The flag identifies whether or not the value of the variable resulted from the use of an imputation method and identifies the imputation method used. For more information on the implementation of the imputation procedures, see Potter and Diaz Tena (2003).

VI. DERIVING APPROPRIATE VARIANCE ESTIMATES

The NSCF used a complex sampling design, which in turn requires that much care should be used when preparing variance estimates. The sampling variance of an estimate derived from survey data for a statistic (such as a total, a mean or proportion, or a regression coefficient) measures the random variation among estimates of the same statistic computed over repeated implementation of the same sample design with the same sample size on the same population. The sampling variance is a function of the population characteristics, the form of the statistic, and the nature of the sampling design. The two general forms of statistics are linear combinations of the survey data (e.g., a total) and nonlinear combinations of the survey data, which include the ratio of two estimates (e.g., a mean or a proportion in which both the numerator and the denominator are estimated) and more complex combinations such as regression coefficients. For linear estimates with simple sample designs (such as a stratified or unstratified simple random sample) or complex designs (such as stratified multi-stage designs), explicit equations are available to compute the sampling variance. For the more common nonlinear estimates with simple or complex sample designs, explicit equations are not generally available and various approximations or computational algorithms are used to provide an essentially unbiased estimate of the sampling variance.

The NSCF sample design involved stratification and unequal probabilities of selection. Variance estimates calculated from NSCF data must incorporate the sample design features in order to obtain the correct estimate. Standard statistical packages used for data analysis, such as SAS and SPSS, are not appropriate for the NSCF design because their assumptions are of independent, identically distributed observations, or simple random sampling with replacement. Although the simple random sample (SRS) variance may approximate the true sampling variance

for some surveys, it is likely to substantially underestimate the sampling variance with a design as complex as the NSCF design. Complex sample designs have led to the development of a variety of software packages that require the user to identify essential design variables such as strata, clusters, and weights. [10]

Sampling variance estimators for complex sample designs take on two primary forms: the procedures based on the Taylor series linearization of the nonlinear estimator using explicit sampling variance equations and the procedures based on forming pseudo-replications [11] of the sample. Within the class of pseudo-replication procedures, the balanced repeated replication (BRR) procedure, the jackknife procedure, and the bootstrap procedure are most widely used and discussed (Wolter 1985). The discussion here will be limited to the Taylor series linearization procedure and the BRR procedures because they are more generally available in survey data analysis software.

A. TAYLOR SERIES LINEARIZATION PROCEDURE

The Taylor series linearization procedure is based on classical statistical method in which a nonlinear statistic can be approximated by a linear combination of the components within the statistic. The accuracy of the approximation depends on the sample size and the complexity of the statistic. For most commonly used nonlinear statistics (such as ratios, means, proportions, and regression coefficients), the linearized form has been developed and has good statistical

---

[10] An Internet site, created with the encouragement of the Section on Survey Research Methods of the American Statistical Association, is now available that reviews software for variance estimation from complex surveys—http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. The site lists software packages available for personal computers and provides direct links to the home pages of these packages. The site also contains articles that provide general information about variance estimation and links to articles that compare features of the software packages.

[11] Pseudo-replications are restricted or random subsamples of a specific survey sample, as opposed to true replications of the sampling design, which entails the selection of multiple independent samples using the same sampling design.

properties. Once a linearized form of an estimate is developed, the explicit equations for linear estimates can be used to estimate the sampling variance. Because the explicit equations can be used, the sampling variance can be estimated using many of the sampling design's features (e.g., finite population corrections, stratification, multiple stages of selection, and unequal selection rates within strata). This is the basic variance estimation procedure used in SUDAAN, Stata, and other software packages to accommodate many simple and complex sampling designs. To be able to calculate the variance, sample design information (such as stratum, analysis weight etc.) is needed for each sample unit.

B. BALANCED REPEATED REPLICATION PROCEDURE

The balanced repeated replication (BRR) procedure is designed for use with stratified multi-stage sample designs in which two primary sampling units are selected with replacement in each stratum. The full sample of primary sampling units is divided into equal-sized half-samples (pseudo-replicates), and the sampling variance is estimated by computing the variation among the survey estimates calculated for each half-sample. The process for forming the half-samples is constrained to ensure a "balance" among the half-samples (Wolter 1985). The BRR procedure was developed by the Census Bureau to estimate sampling variances before the availability of sophisticated high-speed computers for large national surveys. For some estimates for small subpopulations, the BRR procedure cannot compute correct estimates of the sampling variances. To account for this, a modified BRR procedure (Fay's method) is commonly used in which the full sample is used with differential weighting of the half-samples (Judkins 1990).

The BRR procedure is not directly appropriate and adaptations are required to produce unbiased sampling variance estimates for sample designs that use simple stratified random samples, without-replacement sample selection with high sampling rates, or certainty selection of primary sampling units (Rao and Shao 1996; Rao and Shao 1999). In addition, BRR, like other

pseudo-replication methods, requires an initial expenditure of effort to form the replicates, compute a separate set of weights for each replicate, and apply all the nonresponse and poststratification adjustments independently to each replicate. On the other hand, the BRR approach does not require the development of a linearized form of the estimator, so sampling variances can be computed for some forms of complex nonlinear estimates or non-smooth estimators that either cannot be or have not been incorporated in software using the Taylor series linearization procedure. An advantage of replication is its ease of use at the analysis stage because the same estimation procedure is used on all replicates and the full sample, and the actual variance computation is readily computed. The procedure can be applied to most statistics as well as to subgroups. Another advantage is that the procedure accounts for adjustments in weighting the data. By developing weighting adjustments for each replicate, the full effect of the adjustments, such as for nonresponse and poststratification, can be accounted for in the calculation of sampling variances. Software for replication methods requires either replicate weights or sample design information, including the sample weight and stratification information. WesVar (Westat, Inc. see http://www.westat.com/wesvar/) is a popular software program that can compute sampling errors using replication methods.

C. VARIANCE ESTIMATION PROCEDURES AND THE NSCF DESIGN

MPR developed the variance estimation specifications necessary for the Taylor series linearization procedure (PseudoStrata and PseudoPSU). In addition, because of interest in using the BRR procedure for some analyses, 72 pseudo-replicates and the appropriate adjusted and post-stratified weights for each of the pseudo-replicates were computed. The BRR pseudo-replicate weights are variables BRR_WT1, BRR_WT2…BRR_WT72. Appendix D contains example code for both the Taylor Series linearization procedure and the BRR procedure using

the survey data analysis software SUDAAN (Research Triangle Institute 2001). For more information on the variance estimation procedures, see Potter and Diaz Tena (2003).

VII. FILE DETAILS

This chapter provides and overview of the data files, weight variables, and documentation.

A. DATA FILES

As noted earlier, there are two versions of the NSCF data – a Restricted Use File and a Public Use version. Identifiers that could be used to directly identify survey respondents, such and name and social security number, have been omitted from both versions of the data. The Public Use version has been subjected to more aggressive data masking than the Restricted Use File to minimize the likelihood of indirect identification of respondents.[12] This additional masking further protects the confidentiality of survey respondents while simultaneously allowing the use of NSCF survey data by the general research community with a minimum of restrictions.[13] Table VII.1 provides an overview of the two files. The data are available as SAS "sas7bdat" format datasets.

[12] Indirect identification refers to the use of an individual's characteristics, such as age, sex, or geographic location, to identify the survey respondent. Direct identification refers to the use of unique attributes, such as name or SSN, to identify the respondent.

[13] Refer to Chapter V for a discussion of the masking procedures that were employed and a description of the differences between the Public and Restricted versions of the file.

Table VII.1

NSCF DATA FILE DESCRIPTIONS

Version	File Name	Number of Records	Number of Variables
Restricted	Nscffinalfile092003.sasb7dat	9242	2995
Public	NSCF_PUF_122003.sasb7dat	9242	1092

B. WEIGHT VARIABLES

A weight variable, WgtFinal, is provided on both data files. Use of this variable allows estimates of SSA's national analytic populations and subpopulations described in the sample design chapter of this report (Chapter II). This weight should be used when performing any analyses. Due to the design of the NSCF, and the subsequent variation of weights within sampling strata, the use of unweighted, rather than weighted, records will provide incorrect analytic results.

A set of replicate weights, BRR_wt1 through BRR_wt72 are provided to enable the user to use the Balanced Repeated Replication (BRR) variance estimation technique for variance estimation. Refer to Appendix D for example SUDAAN programs that show the proper use of the weight variables.

C. CODEBOOKS

To aid the user, MPR developed codebooks for the restricted and public use data files. The codebooks, available as electronic Adobe Acrobat pdf files, include extensive documentation for each variable including questionnaire text, constructed variable code, user notes, and frequency

information, as appropriate. The codebooks are included as Appendices C, E and F to the User's Manual.

NSCF Restricted File Codebook

The *NSCF Restricted File Codebook* reflects the contents of the final restricted data file, nscffinalfile092003.sas7bdat. For ease of use, the *NSCF Restricted File Codebook* has been divided into two volumes. *Volume I: The Administrative Codebook* includes sampling and weighting variables. The administrative codebook includes all variables drawn from the SSA administrative files (Appendix E). The Volume *II: The Questionnaire Codebook* includes questionnaire variables, constructed variables, imputed variables, and associated imputation flag variables (Appendix C).

NSCF Public Use File Codebook

The *NSCF Public Use File Codebook* reflects the contents of the NSCF Public Use Data File, NSCF _ PUF_122003.sas7bdat. Due to the reduced volume of the public use file, the public use codebook contains both administrative and questionnaire variables in one file. The NSCF Public Use File Codebook is included as Appendix F.

Codebook Format

The restricted and public use codebooks each follow a similar format as described in Table VI.2:

TABLE VII.2
CODEBOOK FIELDS

Field	Field Description	Notes
Variable Name	Variable name in the dataset	Applicable for all variable types. See Chapter V for a description of the variable naming scheme.
Question Number	Hardcopy question corresponding to the variable	Applicable for questionnaire variables only. Because administrative variables, constructed variables, and imputation flags were not asked of respondents, they will not have a question number.
Sample	Total number of respondents from which data are available	Applicable for all variables
Path(s)	Respondent paths from which data are available	Applicable for questionnaire variables only. Administrative, constructed, imputation flag variables, public use, and public use imputation flag variables will not have pathing information. See Table III.I for a description of the paths.
Position	Starting column position of data associated with the variable in the dataset	Applicable for all variables. Position information for each variable is particular to the file in use. Variables may have different position information on the restricted file than on the public use file.
Width	Width (in characters) of data associated with the variable	Applicable for all variables
Type	Type of data: numerical (indicated by num) or character (indicated by char)	Applicable for all variables

Field	Field Description	Notes
Question Source	Source from which the question was taken	Applicable for questionnaire variables only. Administrative, constructed, imputation flag variables, public use, and public use imputation flag variables will not have a question source. See Table III.2 for further information on question sources.
Question Text	English questionnaire text	Applicable for questionnaire variables only. Administrative, constructed, imputation flag variables, public use, and public use imputation flag variables will usually not include question text. However, due to the extensive constructed variable additions in Section K and the deletion of original variables, questionnaire text is included with these constructed variables. Spanish questionnaire text can be viewed in the NSCF questionnaire files.
User Notes	This field provides additional information about data problems, the development of constructed variables, etc.	Included as needed
Constructed Variable Code	This field contains the SAS code used to construct the variable	Included for constructed, public use and public use imputation flag variables
Frequency	Total number of respondents in a particular response category	Applicable for all variables. For ease of use, frequency, value, and label details are suppressed for certain variables including MPRID, the unique identifier and the weight variables.
Value	Numerical value in the dataset associated with a particular response category	Applicable for all variables. See Table V.2 for a description of value labels
Label	Label associated with a particular response category	Applicable for all variables

Using the Codebook files

The NSCF codebook files are available as pdf files, which requires Adobe Acrobat reader for viewing. “Bookmarks” have been inserted into the codebook files to direct the user to specific questionnaire sections or other sections of interest in the codebook. Specific variables of interest can also be located using the FIND function (Ctrl +F), although this function is less helpful for certain variables in the restricted file because of the inclusion of the variables names in the constructed variable code. The most direct method of locating a variable of interest is to navigate (using bookmarks) to the variable’s questionnaire section and then use the FIND function to locate the specific variables.

REFERENCES

Gillcrist, J., S. Mitchell, and D. Kasprzyk. "Report on Data Quality in the National Survey of SSI Children and Families." Washington, DC: Mathematica Policy Research, 2004.

Ireys, H., D. Kasprzyk, A. Takyi, and J. Gillcrist. "NSCF Data, Characteristics of the SSI Child Population: A Comparison Between the NCSF and Three National Surveys." Washington, DC: Mathematica Policy Research, 2004.

Judkins, D.R. "Fay's Method for Variance Estimation." *Journal of Official Statistics*, vol. 6, no. 3, 1990, pp. 223-239.

Mitchell, S., C. Lamothe-Galette, and F. Potter. "Survey Response Incentives for a Low-Income Population: What Works?" Issue Brief #2. Washington, DC: Mathematica Policy Research, Inc., November 2003.

Potter, F. "Report on Revised Sampling Design: National Survey of SSI Children and Families." Washington, DC: Mathematica Policy Research, July 2000.

Potter, F.J., and S. Mitchell. "Report on Sampling Design and Estimated Survey Costs: Evaluation of the Effects of the 1996 Welfare Reform Legislation on Children with Disabilities: Survey Design and OMB Clearance Package." Washington, DC: Mathematica Policy Research, 2004.

Potter, F., and N. Diaz-Tena. "Weighting, Nonresponse Adjustments, and Imputation: National Survey of SSI Children and Families." Washington, DC: Mathematica Policy Research, July 2003.

Rao J.N.K., and J. Shao. "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling." *Journal of the American Statistical Association*, vol. 91, 1996, pp. 343-348.

Rao J.N.K., and J. Shao. "Modified Balanced Repeated Replication for Complex Survey Data." *Biometrika*, vol. 86, 1999, pp. 403-415.

Research Triangle Institute. *SUDAAN User's Manual, Release 8.0.* Research Triangle Park, NC: Research Triangle Institute, 2001.

Wolter, K.M. *Introduction to Variance Estimation.* New York: Springer-Verlag, 1985.

LIST OF APPENDICES